PROGRESS IN STANDARDIZING THE MEASUREMENT OF COMPOSITION[1]

ERNEST C. NOYES
Fifth Avenue High School, Pittsburgh, Pa.

I wish to call attention to a recent investigation which has for its purpose, "to establish standards of composition that will make it possible to compare the work done in one school with that done elsewhere and to make it difficult for mere opinion to control so much of our schoolroom practice." I refer to the attempt of Professor Thorndike and Mr. Hillegas of Columbia University to construct a concrete scale of measurement for composition.

In nearly every form of human effort except teaching, the efficiency of different methods of procedure can be and is tested by results. That school of medicine whose practitioners cure the greatest proportion of their patients quickly becomes the most popular. Comparison of the results gained by automobile delivery with those gained by horse and wagon is driving the horse from our streets. Even in the humble field of poultry-raising comparison of the number of eggs laid in a month by the same number of hens under different systems of feeding has proved what kind of food will produce the greatest results in eggs. It is only in teaching, and this is especially true of the teaching of English, that when our favorite theory is subjected to the test of pragmatism, Does it work? we must answer, "We don't know positively, but we hope it works, or we think it ought to work."

Now just as the study of methods of business has produced the new application of science to industry known as scientific management, so the demand of the age for the measurement of results is bringing forth a new science of education based upon exact measurement and judgment by ascertained facts. As Professor Thorndike has said, "It is fruitless to keep only the debit account of time and

[1] Read before the Joint Conference on English at the National Education Association in Chicago, July 11, 1912.

money expended, of courses of study and methods of teaching, if we leave the credit account of results achieved, the products of education, vague and uncertain." The need of measuring our results requires no demonstration; but the keeping of an accurate account presents a problem of peculiar difficulty. If education has lagged behind business in testing theories by results, it has been because there have existed no adequate means of gauging skill in solving arithmetical problems, of measuring knowledge of Latin, or of estimating ability to write English.

To be sure, we have for decades attempted to measure composition by percentages or letters, or by vague adjectives like good, fair, and excellent. But can any teacher tell just when a composition ceases to be worth 90 and is worth only 85, or when it rises to 95? And will any teacher assert that the difference between 50 and 60 is the same as that between 90 and 100? Yet in any valid scale ten units should represent the same step in whatever part of the scale they be taken. If we abandon the percentage system in favor of letters and make A stand for excellent, B for good, C fair, and so on, we find it equally hard to draw a line and say that all paragraphs on one side of this line are fair and that all on the other are poor. In addition, we are compelled with the letter system to give the same mark to compositions varying greatly in merit. This has caused many schools to add plus and minus signs to the letters; but who can determine the point at which a composition falls below B minus in merit and yet is worth C plus? Furthermore, whatever system is used, all of us know that teachers working independently will scarcely ever grade a set of compositions alike.

In estimating the merit of specimens of composition, teachers probably show greater differences in judgment than they exhibit in grading pupils' performances in any other subject. Though the extraordinary complexity of the facts to be observed and estimated in forming any judgment, together with the variable emphasis placed by individuals upon vocabulary, sentences, paragraphing, thought, etc., makes uniformity of judgment difficult, it is not impossible. Differences in the rating of the same paper exist not because teachers disagree about what constitutes

merit in composition, for on essentials there is in theory substantial agreement. The chief difficulty is the vagueness with which standards have been defined. When we say of a boy's composition that the sentences, paragraphing, and progress of thought are *good*, we are as inaccurate as we are when we say it was *hot* here yesterday. If we say the thermometer stood at 90° here yesterday, we have measured the heat by a scientific standard which has the same meaning in Boston as in Chicago and by which we can compare the heat in one place with that in another. On the other hand, if we say that the composition is worth 90 per cent, though we have measured it after a fashion, the standard does not have the same meaning in Boston as in Chicago. It is not the same with any two teachers, and is more worthless for purposes of comparison than 90° Fahrenheit and 80° Centigrade would be; for the two temperature scales can be reduced to a common basis, but no two teachers' percentages have any common basis. Our present methods of measuring compositions are controlled too much by personal opinion, which varies with the individual. What is wanted is a clear-cut, concrete standard of measurement which will mean the same thing to all people in all places and is not dependent upon the opinion of any individual.

This want Professor Thorndike and Mr. Hillegas[1] have attempted to supply by the construction of a series of samples of composition with which specimens to be measured can be compared. This series of samples, each differing from the preceding by an equal degree of merit, makes a uniform scale of measurement *applicable universally*. In applying this scale, a specimen of composition to be valued is compared with one sample after another till a sample is found precisely equal in merit, just as in determining the note sounded by a tuning-fork we had found we might compare it with one note after another of the musical scale sounded on a piano till we could fix its proper place.

Though at first thought it may seem impossible to construct such a scale for composition, the means by which it has been done is as simple as it is sound; namely, the opinion of a considerable

[1] A full account of his scale is given by Mr. Hillegas in the *Teachers College Record* for September 1912.

number of competent judges. Certainly, the test that has determined what books are classics and that lies behind the principle of good use in language is the best that could be devised for determining a scale of measurement for composition. Mr. Hillegas' scale represents the rankings of several hundred carefully selected judges, who were asked to arrange a large number of samples of young peoples' composition in order of merit. From their rankings of this large set, the scale of ten samples, ranging in value by equal steps from 0 to 937 units was derived by applying the theory that, "Differences equally often noticed are equal, unless the differences are either always or never noticed." That is, if 75 per cent of the judges noticed that sample A was better than sample B and the same number noticed that B was better than C, it was assumed that the difference in merit between A and B was exactly the same as that between B and C; or that the difference between A and C was exactly twice the difference between B and C, or B and A. In fact, when two samples were found such that 75 per cent of the judges agreed in calling one better than the other, the difference in merit between them is just the difference used as the unit in this scale. The zero point is shown by an artificial sample produced by an adult who tried to write very poor English.

Of course, such a scale is open to some objections, notably that the samples it contains, which vary in length from five to ten printed lines, are too short to test sustained power or skill in arrangement of the larger units of composition. Some may say that judgment of composition by the use of such a scale is too mechanical. Still, what is essentially the same method, comparison with samples of fixed value, was recommended by no less sensitive a critic than Matthew Arnold when he advised the committing to memory of fine passages of poetry that they might serve as touchstones by which to test the value of unfamiliar verse.

This scale marks a great advance toward precision in the rating of composition. It will enable any teacher to correct his individual judgment by reference to the combined opinions of many good judges. It will afford him a ready means of determining the relative progress of pupils under different systems of instruction so that when asked, "Does this theory work?" he can say, "Yes"

or "No" with conviction. Even mental standards will gain in precision as teachers become accustomed to a uniform standard. Composition equivalent to six on Hillegas' scale will be a much more accurate description than fair; just as our familiarity with the thermometer has made fifty degrees a more precise description than chilly.

It is to supervisors, however, that this scale will be of the most value. By it they will be able to compare classes of the same grade in different schools, in different cities, or under different teachers. Such a scheme for measurement will also afford accurate definitions of the standards required for promotion. Employers will find in it an accurate means of defining the degree of excellence expected in an applicant for a position. Even the pupil may profit by studying such a scale, for it will show him just what he is expected to accomplish and enable him to measure his own progress.

In short, Mr. Hillegas has shown that it is possible to standardize our measurement of composition and to make it precise. Soon we shall think it as ridiculous to measure compositions by excellent, good, and fair, as to measure distance by units like as far as a man can jump. This investigation signifies the dawn of the day when the incapable teacher shall be unable to take refuge in the belief that no one can measure the results of his work; when the thoughtful teacher may be able to tell by accurate tests whether his theories are true or false; and when all teachers may feel that their gradings are just and uniform because based on a definite, fixed standard.